AVCD: Mitigating Hallucinations in Audio-Visual Large Language Models through Contrastive Decoding

Chaeyoung Jung*, Youngjoon Jang*, Joon Son Chung

KAIST

*Equally contributed

Code:

NEURAL INFORMATION PROCESSING SYSTEMS

Background

(a) Hallucination in Audio-Visual Large Language Models (AV-LLMs)

Question
"How many sounding acoustic guitar in the video?"

Base	VCD	AVCD
"One"	"One"	"Zero"

- Hallucinations arise from complex cross-modal interactions
- AV-LLMs struggle when modalities conflict (*e.g.,* Video shows a guitar, audio plays a violin → model answers "one sounding acoustic guitar.")
- Trimodal reasoning is significantly harder than VL (vision-language).

(b) Contrastive Decoding (CD) for mitigating hallucination in Vision-Language Models (VLMs)

Original Visual Inputs v

Distorted Visual Inputs $\neg v$

VLM → $\text{logit}(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^l)$

VLM → $\text{logit}(\mathbf{y}|\mathbf{x}^{\neg v}, \mathbf{x}^l)$

$$\text{logit} = (1+\alpha)\text{logit}(\mathbf{y}|\mathbf{x}^v, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{y}|\mathbf{x}^{\neg v}, \mathbf{x}^l) \quad (2)$$

- CD compares the original logits with logits from distorted inputs
- In VLMs, CD perturbs the visual modality to create distorted inputs
- Traditional CD uses only one negative (corrupted) instance

(c) Tri modal interactions in AV-LLMs

- Hallucinations in AV-LLMs arise across all levels: Unimodal, Bimodal, Trimodal
- Trimodal reasoning introduces new challenges
- Traditional CD uses fixed modality corruption and not designed for trimodal models.

↓

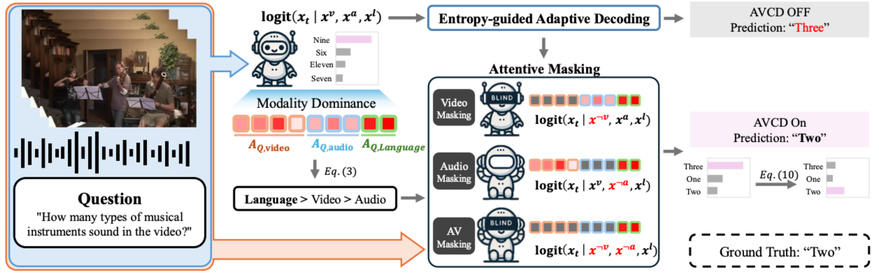
Propose Audio-Visual Contrastive Decoding (AVCD):

1. Handle up to 3 modalities

2. Adaptive hallucination mitigation

3. Improves inference efficiency

Method: AVCD



1. Measuring modality dominance

$$D_M^j = \sum_{i \in M} A_{Q_K^j, i},\ D_M = \frac{1}{J}\sum_{j=1}^{J} D_M^j, (3)$$

- Dominance score D_M, D_M^j represents **the dominance score of modality M.**
- $A_{Q_K^j, i}$ is the **attention weight assigned by the last query token** to i-th key token.
- M represents the set of indices corresponding to video, audio, or language tokens.

2. Attention masking via zeroing out

- We mask **less dominant modalities** (e.g., when language is dominant, video and audio are masked).
- Masking uses an **attention-based threshold** (the **top P% of the mean stacked** A_{Q_K} across layers).

3. A reformulated CD for trimodal integration

- Modeling the probabilities in Eq. **(2)** when language is dominant as follows:

$$p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \sum_{a' \in \{a, \neg a\}} \frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{a'}, \mathbf{x}^l)\ (4),\ p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^l) = \sum_{a' \in \{a, \neg a\}} \frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{a'}, \mathbf{x}^l)\ (5)$$

- To convert the distribution into logit form, we apply the logarithm to Eq. **(4)** as follows:

$$\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \log\left(\frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \frac{1}{2} p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right). (6)$$

- Directly adding probabilities inside the logarithm is undesirable, therefore we use Taylor expansion

$$\log\left(\frac{A+B}{2}\right) = \frac{\log(A)+\log(B)}{2}. (7)$$

- Accordingly, we can approximate $\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l)$ as follows:

$$\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) = \frac{1}{2}\left(\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)\right). (8)$$

- Similarly, by applying the logarithm to Eq. **(5)** and substituting the results into Eq. **(2)**, we derive the extended CD for the visual modality:

$$\text{logit}_v \propto (1+\alpha^v)\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^l) - \alpha^v \log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^l) \propto (1+\alpha^v)(\log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l)) - \alpha^v(\log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) + \log p(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l)) \quad (9)$$

- Address both video and audio, we sum logits and final CD:

$$\text{logit}_{\text{AVCD}} \propto (2+\alpha^v+\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) + (1-\alpha^v+\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) + (1+\alpha^v-\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l) - (\alpha^v+\alpha^a)\text{logit}(\mathbf{y}_t|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l) \quad (10)$$

Results

Quantitative results

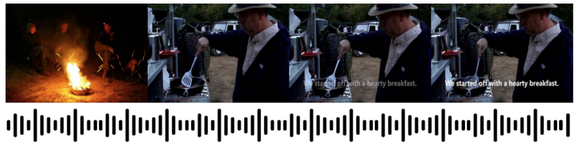
Table 3: **Results on video-LLMs.** AVCD surpasses both Base and VCD across all experiments.

Model	Decoding	Datasets	
		MSVD-QA [62]	ActivityNet-QA [67]
		Acc. (%) ↑	Acc. (%) ↑
VideoLLaMA2 [11]	*Base*	74.43±0.31	47.19±0.55
	VCD [28]	71.30±0.57	45.65±0.04
	AVCD	**75.20**±0.42	**48.22**±0.04
Video-LLaVA [35]	*Base*	70.20±0.20	47.48±0.02
	VCD [28]	71.80±0.25	47.25±0.02
	AVCD	**72.16**±0.24	**48.03**±0.14

Table 2: **Results on AV-LLMs.** We evaluate two representative AV-LLMs across three datasets. For decoding, we compare the original model's decoding (*Base*), VCD [28] extended with audio via Eq. (2), and VCD*, which incorporates audio using our proposed formulation in Eq. (10) along with adaptive dominant modality recognition. AVCD consistently outperforms all other decoding methods across the benchmarks, demonstrating the effectiveness of both our trimodal CD formulation.

Model	Decoding	Datasets		
		MUSIC-AVQA [29]	AVHBench [51]	AVHBench-Cap
		Acc. (%) ↑	Acc. (%) ↑	Score ↑
VideoLLaMA2 [11]	*Base*	81.30±0.09	70.52	2.84±0.01
	VCD [28]	77.66±0.03	65.18	2.86±0.01
	VCD*	81.49±0.03	69.18	3.00±0.01
	AVCD	**81.58**±0.03	**72.15**	**3.03**±0.01
video-SALMONN [50]	*Base*	48.50±0.06	58.19	1.83±0.01
	VCD [28]	41.57±0.08	60.61	2.41±0.01
	VCD*	49.00±0.11	60.66	2.44±0.01
	AVCD	**49.73**±0.06	**62.18**	**2.47**±0.02

Qualitative results



Prompt: Describe the given video and audio in detail.

*Visual mismatch

Base
Under the soft glow of a campfire, a group of friends gathers around a table, immersed in conversation and camaraderie. The warmth of the fire dances in their eyes as they share stories and laughter, punctuating the night air with the occasional crackle and pop of the burning wood. As the evening deepens, the flickering flames cast an ever-changing light on their faces, creating an atmosphere of cozy comfort and shared joy. In the distance, the quiet hum of the wilderness echoes off the surrounding foliage, adding to the serene ambiance of this nocturnal gathering.
❌ Fail to describe the scene change and audio information

VCD
Under the calm of the night sky, **a small group gathers around a crackling campfire**, engaged in conversation and the simple joy of being together. The warmth of the flames dances off their faces, as they share stories and laughter, the embers casting a soft glow on their smiling faces. As the night grows late, the fire's light becomes a beacon, drawing in the quiet stillness of the evening, symbolizing a moment of camaraderie and connection under the vast, dark canopy above.
❌ Fail to describe the scene change and audio information

AVCD
Under the calm of a night sky, **a small group gathers around a crackling campfire**, immersed in conversation and the shared camaraderie of an outdoor adventure. The warmth of the flames dances off their faces, accentuating the outlines of their relaxed postures as they lounge **in their camping chairs**. The rhythmic popping of the fire provides a soothing backdrop to the soft murmur of their conversation.
As the night deepens, **one of the group takes on the role of cook,** reviving the fire with a **hearty breakfast**. Amidst **the sizzling and clinking of utensils**, the aroma of cooking wafts through the air, mingling with the crisp, earthy scent of the campsite. The cook's skilled hands work methodically, **flipping and stirring the food in a black cast-iron skillet**, ensuring that every morsel is cooked to perfection.

Further analysis

Table 4: **Ablation on CD with Eq. (2) and Eq. (11).** AVCD effectively extends existing CD to trimodal configurations.

Decoding	Masked Modality			Acc ↑
	A	V	AV	
Base				70.52
w/ Eq. 2	✓			71.88
w/ Eq. 2		✓		70.16
w/ Eq. 2			✓	72.07
w/ Eq. 11	✓	✓	✓	70.94
AVCD	✓	✓	✓	**72.15**

$$\text{logit} = (1+3\alpha)\text{logit}(\mathbf{x}|\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^{\neg v}, \mathbf{x}^a, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^v, \mathbf{x}^{\neg a}, \mathbf{x}^l) - \alpha\text{logit}(\mathbf{x}|\mathbf{x}^{\neg v}, \mathbf{x}^{\neg a}, \mathbf{x}^l). (11)$$

Table 5: **Ablation study on dominant modality recognition.** We compare fixed dominant modality settings with our adaptive recognition strategy. AVCD consistently outperforms static configurations, demonstrating the effectiveness of dynamic modality selection.

Dominant Modality	Dataset	
	AVHBench [51]	MSVD-QA [62]
Audio	68.67	-
Vision	67.79	70.70±0.14
Language	**72.15**	74.20±0.14
Adaptive	**72.15**	**75.20**±0.42

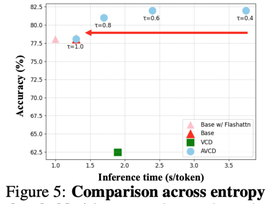


Figure 5: **Comparison across entropy thresholds (τ).** τ controls over the trade-off between inference speed and accuracy. At $\tau = 0.8$, it achieves faster inference than VCD while outperforming *Base* decoding in accuracy.